sciendo INTERNATIONAL JOURNAL ON SMART SENSING AND INTELLIGENT SYSTEMS

Article | DOI: 10.2478/ijssis-2023-0006 Issue 1 | Vol. 16 (2023)

Identification of glucose levels in urine based on classification using *k*-nearest neighbor algorithm method

Anton Yudhana[1,*], Fathiyyah Warsino[1], Son Ali Akbar[1], Fatma Nuraisyah[2] and Ilham Mufandi[1,3]

[1]Department of Electrical Engineering, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Unit IV, Jl. Ringroad Selatan, Kragilan, Tamanan, Banguntapan, Bantul Regency, Special Region of Yogyakarta 55191, Indonesia

[2]Public Health Science Study Program, Faculty of Public Health, Universitas Ahmad Dahlan, Unit III, Special Region of Yogyakarta 55191, Indonesia

[3]Department of Agro-industrial Technology, Faculty of Science and Technology, University of Darussalam Gontor, Ponorogo 63471, Indonesia

*E-mail: eyudhana@ee.uad.ac.id

This paper was edited by Subhas Chandra Mukhopadhyay.

Received for publication: August 15, 2022.

Abstract

Glucose monitoring carried out through the urine testing to make it easier for patients to check their blood sugar without having to physically injure themselves and to prevent external bacteria from entering the body, which happens while using needles. This study aims to classify glucose-containing urine specimens based on diabetes levels by using the K-nearest neighbor method. Classification of urine specimens is achieved by using the Benedict method to produce the color of the urine specimen and the AS7262 sensor to detect the color produced by the specimen. The results showed that the classification of data on urine specimens has an accuracy of 96.33%. Previous studies conducted this experiment using a photodiode sensor and a TCS sensor, which produced red, green, and blue (RGB) colors. For identifying the color of a specimen, the AS7262 sensor can produce six colors (red, green, blue, yellow,violet, and orange) to identify the glucose level.

Keywords

Glucose, urine, AS7262, classification, K-nearest neighbor

1. Introduction

Diabetes mellitus is one of the diseases with the highest number of sufferers in the world. According to the World Health Organization (WHO), as many as 415 million people suffered from this disease in 2019 [1]. Early diagnosis and ongoing control are important for diabetics to ensure a healthy life and to avoid complications and death [2].

Measurement of glucose levels in the urine is carried out to prevent the entry of bacteria into the body through a syringe [3]. Obesity triggers an increase in glucose due to the accumulation of fatty tissue in the body [4]. When glucose levels in the blood are >180 mg/100 mL, glucose is extracted by the kidneys and flushed out through the urine [5,6]. The level of glucose contained in the urine is thereafter determined.

This study designed a prototype to detect glucose in urine using the AS7262 sensor as the main component for detecting the color of urine specimens. These specimens are grouped into five classes, namely, Normal, Positive 1, Positive 2, Positive 3, and Positive 4. Urine produces color changes according to glucose levels based on Benedict's test [7]. The analysis of this research is strengthened using machine learning (ML) with the *K*-nearest neighbor (KNN) algorithm.

Open Access. Published by Sciendo. BY-NC-ND
© 2023 Yudhana et al. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License https://creativecommons.org/licenses/by-nc-nd/4.0/

2. Materials and Methods

2.1. System design

This system is designed by researcher to detect the color of the specimen using the AS7262 sensor based on the intensities of six colors. They are violet, orange, blue, green, yellow, and red. The supporting components are installed in this system including: the DS18B20 sensor to regulate temperature, a heater to process the specimen, van as a coolant to control the temperature when overheated, buzzer and light-emitting diode (LED) as markers of each process, printed circuit board (PCB) as a microcontroller, liquid crystal display (LCD) to show the menu and the result, a button to control the system, and a switch to turn on or turn off the system. Block diagrams of the system design are made to explain the overall concept specifically of the prototype shown in Figure 1.

Figure 1 shows that this prototype uses a digital communication line, while the AS7262 sensor uses the Inter Integrated Circuit System (I2C). Note that the AS7262 and DS18B20 sensors are input. The value generated from DS18B20 is used to set the temperature on the heater to work automatically, and the AS7262 value is used as a dataset in the next process. The wiring diagram is shown in Figure 2.

Figure 2 shows that each pin on the Arduino Nano is connected to each component in order to control the performance of each component through a program using digital communication. The flow of data collection specifically can be seen in Figure 3.

Figure 3 shows that after the initialization process, the temperature is set and the heating process is started. This process produces color in the specimen and is detected using the AS7262 sensor, with the resulting data displayed on the LCD and sent to the serial monitor. The results of the data are used as a dataset to be processed in the next step.

The data generated by the sensor are processed during ML using the KNN algorithm. The data process in ML is shown in Figure 4.

ML is used to extract relevant data [8]. The dataset used contains the results of the AS7262 sensor, which is converted into comma-separated value (csv) form so that it can be read using Python program. The data are processed for classification to determine the glucose levels based on the color of the specimen. After the data are classified, the algorithm begins to predict, and the accuracy of the classification process is determined.

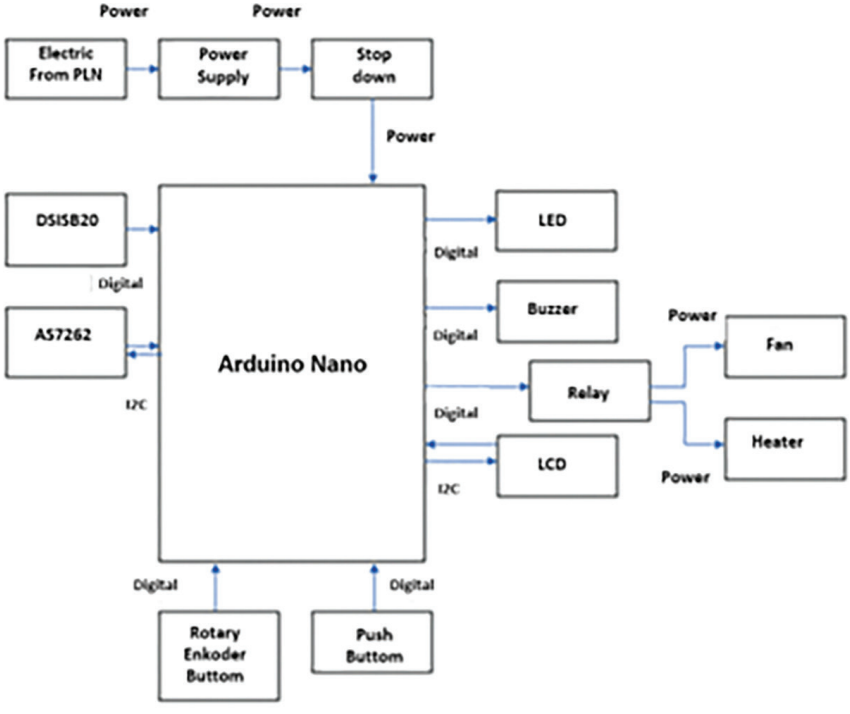


Figure 1: Block diagram. PLN, State Electricity Company; I2C, Inter Integrated Circuit System.

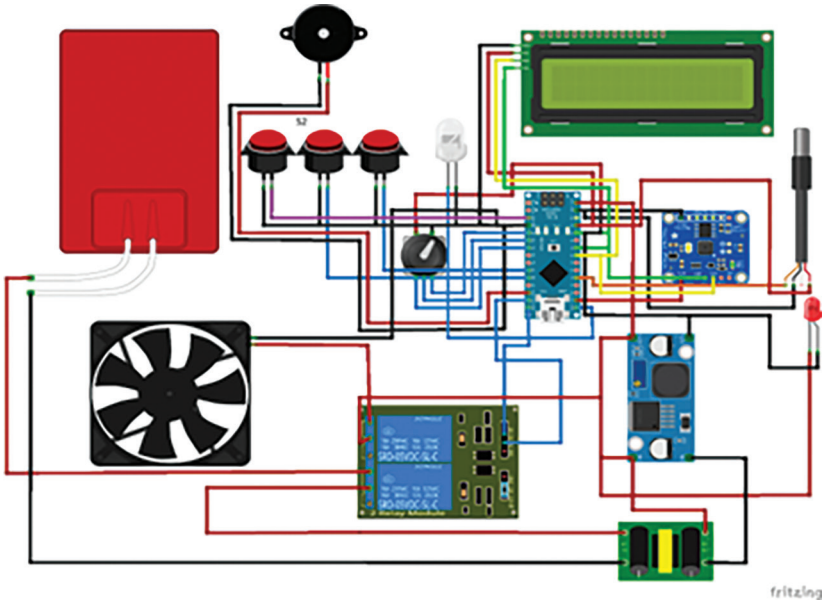


Figure 2: Wiring diagram.

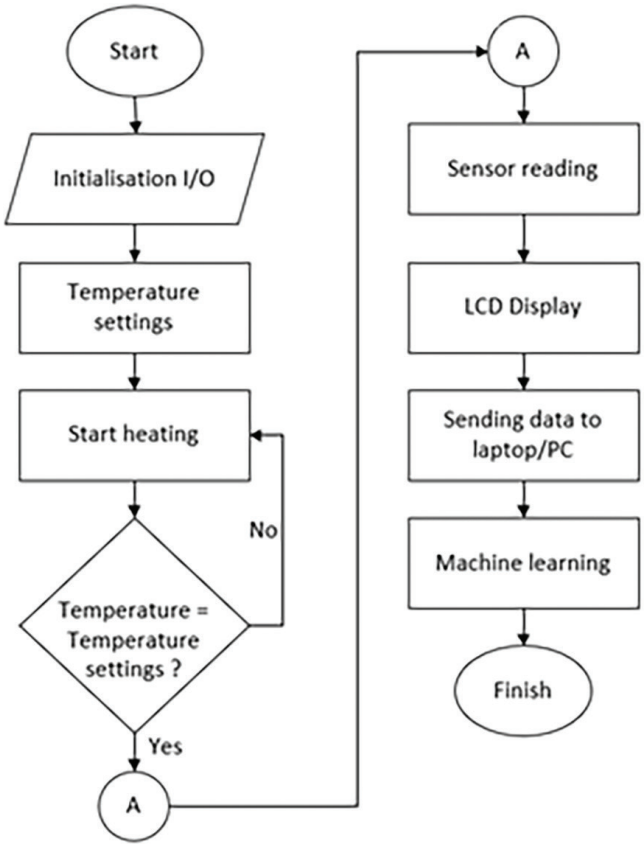


Figure 3: Flowchart.

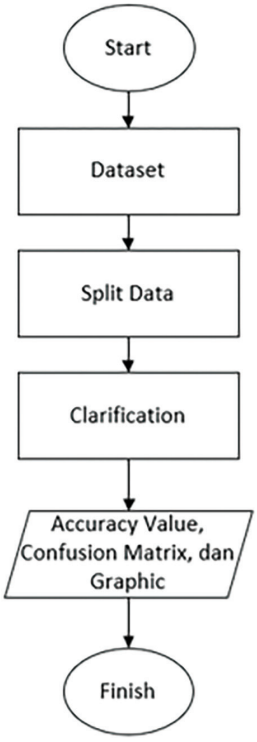


Figure 4: Machine learning flowchart.

2.2. Procedure of collecting data

2.2.1. Sample preparation

The specimens in this study were mixtures of urine processed by Benedict's method to produce a colored precipitate, which was assayed according to the grade of the total glucose level. Benedict's test refers to the provisions contained in previous studies (Pratiwi et al. 2020), with the test using as many as five urine samples with different glucose levels. The difference between such studies and the current study lies in the use of different sensors; previous studies used photodiode sensors to detect colors, while the current study uses AS7262 sensors to detect colors [9]. Benedict's test sample was prepared by using 10 mL of Benedict's solution and 20 mL of urine specimen (containing 0.4 mL, which is equivalent to eight drops). This test is done to detect the glucose content in the urine. The calculation of the concentration is done by using Eq. (1).

$$m = \frac{(\%)\,Solute}{100\%} V\ solution. \quad (1)$$

Here, m is the total of glucose level, '(%) Solute' is the glucose level, and V solution is the total volume of the specimen solution [5].

2.2.2. *K*-Nearest Neighbor Algorithm

The working principle of KNN is to make predictions based on the proximity of the object's characteristics to those of the neighborhood training data closest to the object [10, 11]. This algorithm works by determining the input in the form of training data, testing data, and the value of K [12]. The value of K is the number of nearest neighbors of the data to be classified to determine the class label [13,14].

The advantage of this method is that data are classified based on the nearest neighbor class. The value of "K" in KNN indicates the number of neighbors that are taken into account in determining the class [15]. KNN is widely used in ML because of its simple performance [16]. The visualization of the KNN algorithm is shown in Figure 5.

Figure 5 shows the data, which consist of two classes, namely, the red class and the green class. The test data are indicated by a blue arrow, with a value of $K = 5$. With the number of K values, the test data are classified in the green class based on the majority of the nearest neighbors. A high K-value results in lower accuracy because it can make the boundaries between each classification become more blurred [17]. The calculation of the neighbor distance is done by using the Euclidean equation, which is shown in Eq. (2).

$$euc = \sqrt{\left(\sum_{i=1}^{n} (pi - qi)^2\right)}, \quad (2)$$

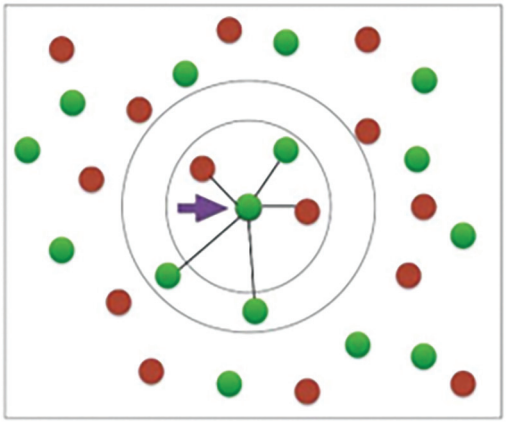


Figure 5: K-nearest neighbor visualization.

where *pi* = sample data, *qi* = test data/testing data, *i* = variable data, and *n* = data dimension [18]. Tests were carried out to determine the level of accuracy of classification of data using KNN. Accuracy is the level of closeness of the predicted value to the actual values of the precision, recall, and *f*-measure [19]. The calculation of accuracy is done as shown in Eq. (3).

$$Accuracy = \frac{Valid\ value}{Total\ data} \times 100\%. \quad (3)$$

Precision is defined as a measure of accuracy to predict all data samples [20]. The precision value is calculated by using Eq. (4).

$$Precision = \frac{TP}{TP + FP}. \quad (4)$$

Here, TP indicates true positive, and FP indicates false positive.

Recall is a completeness measure of the sample scale taken with all significant samples [20]. The recall value is calculated by using Eq. (5).

$$Recall = \frac{TP}{TP + FN}. \quad (5)$$

Here, FN indicates false negative. *F*-measure is the mean or average value between precision and recall [21]. For the calculation of the *F*-measure value, see Eq. (6).

$$F - \text{measure} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (6)$$

Description of the variables:
TP = true positive;
TN = true negative;
FP = false positive;
FN = false negative.

TP indicates the presence of a characteristic, TN indicates the absence of a characteristic, FP indicates the presence of a certain condition, and FN indicates the absence of a certain condition.

3. Results and Discussion

3.1. Sample test

Tests carried out on urine samples containing glucose with Benedict's test resulted in five specimen colors, classified as Normal, Positive 1, Positive 2, Positive 3, and Positive 4. Classes were characterized by grade, color yield, and glucose levels. The results of the characterization can be seen in Table 1.

Based on this classification, it can be observed in Table 2 that the urine specimens in each class have different characteristics. This is caused by the amount of glucose levels contained in the urine. Calculation of glucose concentration is done using Eq. (1).

3.2. Data collection

This process is carried out using a prototype design made with the aim of making it easier for users to retrieve specimen data in terms of time and place. This prototype is made using components that function to help the prototype carry out its functions properly. The prototype in this study is shown in Figure 6.

Figure 6(a) shows the output results for each input displayed on the LCD. Figure 6(b) is the layout of the component series for the data collection process. Figure 6(c) shows the main component AS7262 sensor.

Table 1. Glucose characterization

No.	Urine sample	Glucose level (%)	Glucose rate (g)	Color result	Urine positive rate (mg/dL)
1.	Sample 1	0–0.5	0	Slightly greenish blue and a bit cloudy	Normal
2.	Sample 2	0.5–1	0.2	Yellowish green	Positive 1
3.	Sample 3	1–1.5	0.3	Greenish yellow	Positive 2
4.	Sample 4	2–3.5	0.4	Slightly brownish orange	Positive 3
5.	Sample 5	>3.5	1	Slightly brownish brick red	Positive 4

Table 2. Results of the data from the sensor

Result of specimen color	Sensor color indicator						Class category
	Violet	Blue	Green	Yellow	Orange	Red	
	21.34	23.36	18.68	19.64	14.02	6.66	Normal
	14.93	13.53	24.27	21.48	19.23	6.53	Positive 1
	12.19	10.07	22.63	24.14	21.87	9.05	Positive 2
	11.12	8.21	16.97	21.91	26.02	15.74	Positive 3
	13.68	9.93	22.58	20.77	22.58	10.43	Positive 4

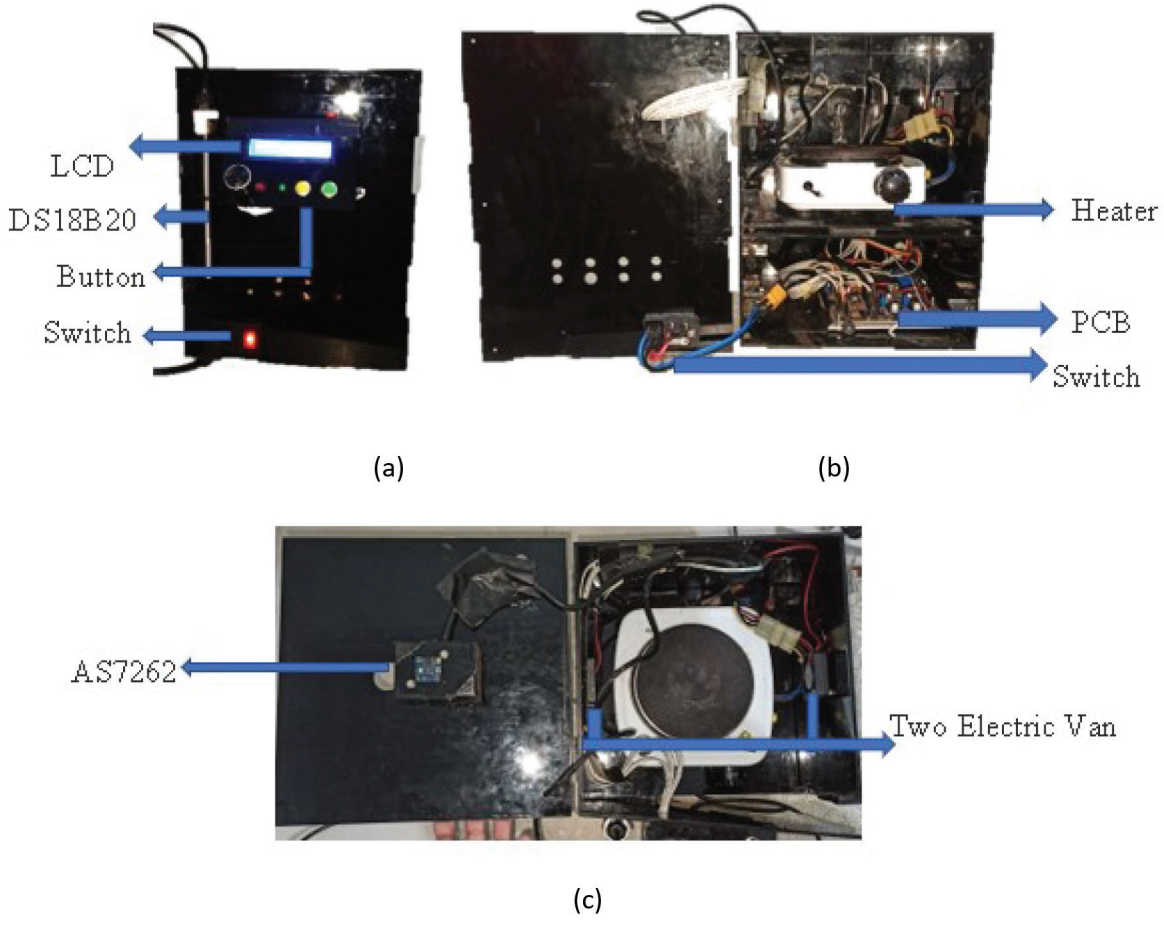


Figure 6: Prototype: (a) front view, (b) inside view, and (c) upper view.

3.3. Data collection

Data testing is carried out to test the accuracy of the sensor by identifying the color intensity of the specimen based on six indicators. Testing this prototype resulted in 1,200 data on each specimen. The total number of data generated by the five specimens is 6,000. The data results are shown in Table 2.

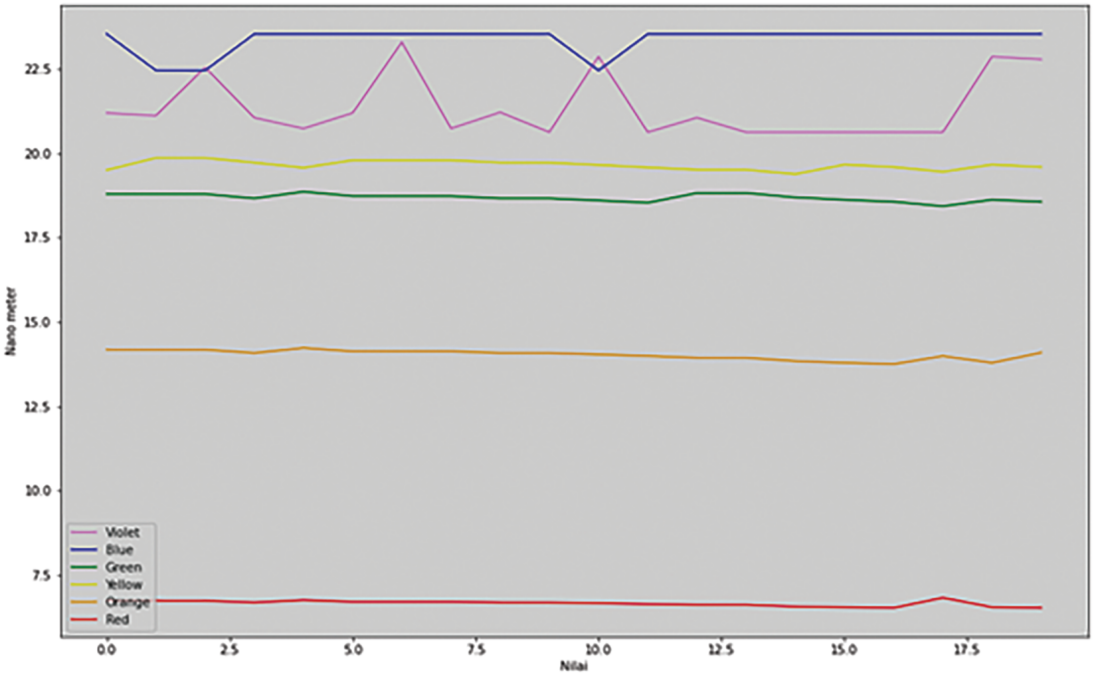
Table 2 shows some of the data results obtained from the AS7262 sensor. The resulting data are quite stable because the sensor produces color intensity data that matches the color of the specimen. The data are displayed in graphical form in Figure 7.

3.2. Classification of glucose levels using KNN

Classification is the process of analyzing the same model on a set and classifying it into different classes [22,23]. The data are classified into five classes based on the glucose levels, with a *K* value of 45; of the 6,000-item dataset, the testing data comprised 30% and the training data comprised 70%. Prediction from the data led to an accuracy of 96.33%. Accuracy is predicted using Eq. (3). The classification is achieved using Eq. (2) based on the distance of each neighbor. The purpose of this process is to obtain a classification of five classes based on six color indicators. The accuracy value is shown in the form of a graph in Figure 8.

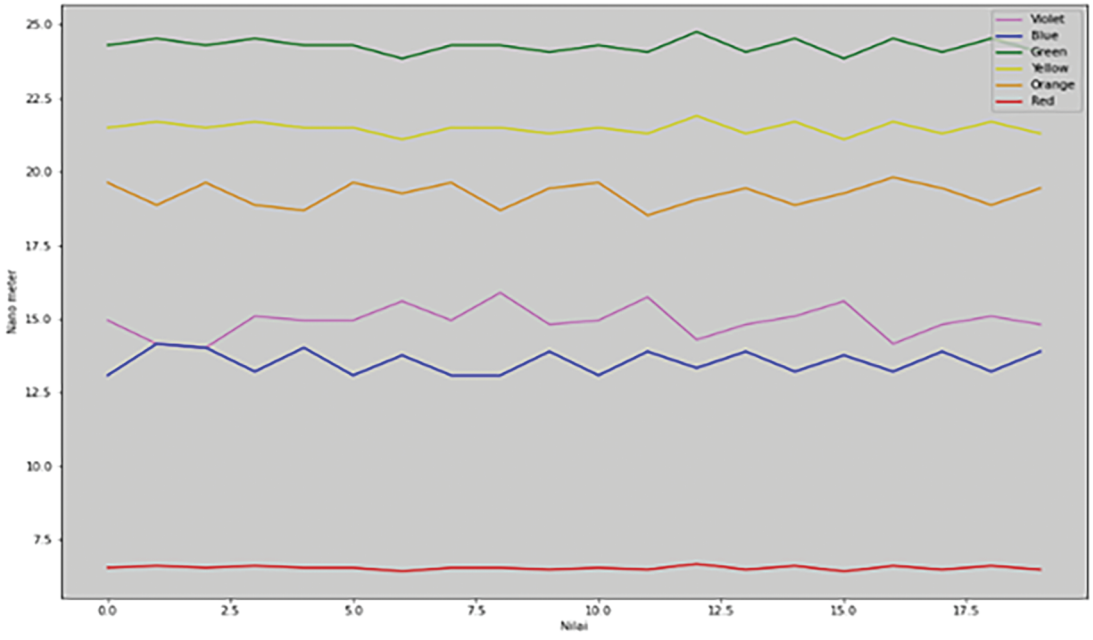
Figure 8 shows that the higher the graph, the more are the error data. The graph moves up and down because of the error data for the value of *K*. The unstable graph is caused by several reasons, including the suitability of the data [24]. The classification process works on the KNN algorithm based on a dataset that is calculated using precision, recall, and *F*-measure. The resulting value for each variable is shown in Figure 9.

Figure 9 shows the results of the predicted values generated from the test data, training data, and the value of *K*. The figure presents the precision, recall, and *f-score* values, which are calculated based on the formulas in Eqs. (5)–(7) with support values in each class. To analyze the classification performance of the KNN algorithm, the confusion matrix shown in Figure 10 can be used.

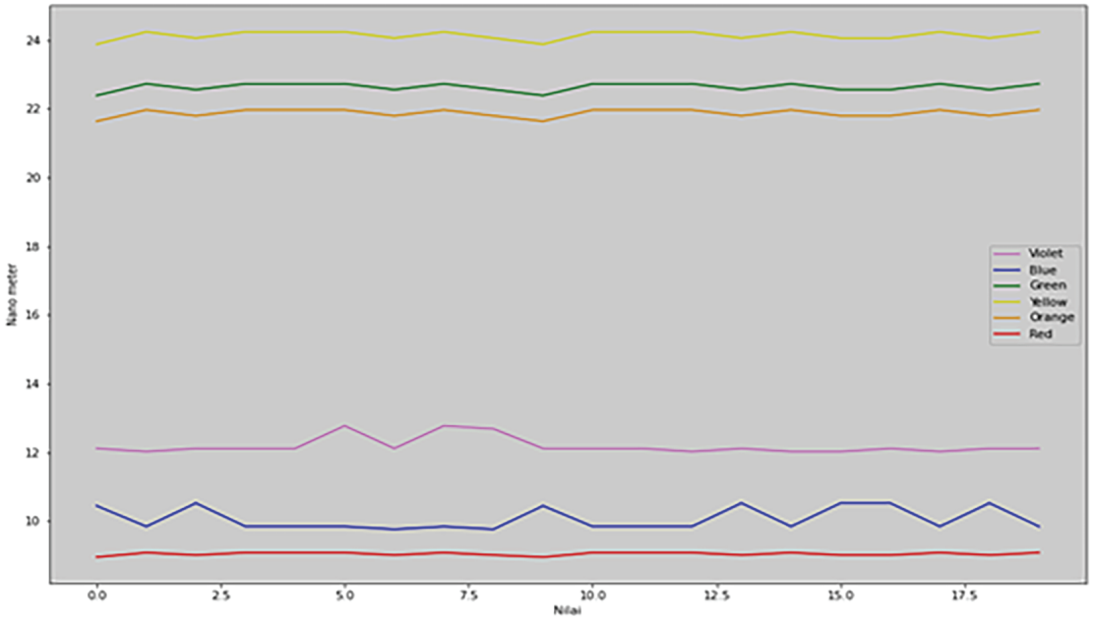


(a)

Figure 7: (Continued).

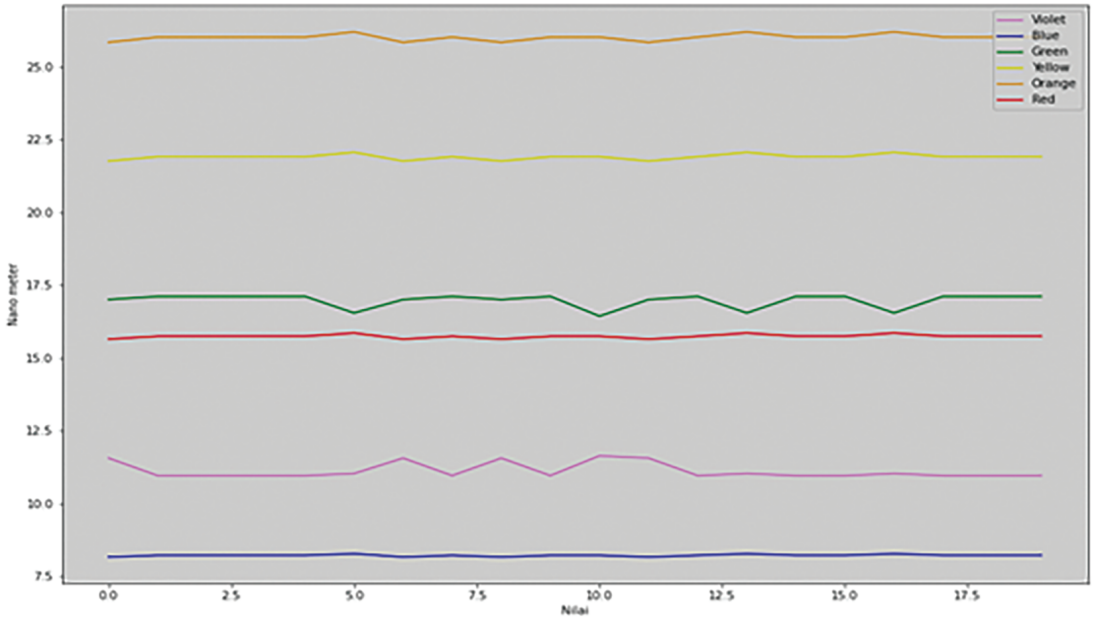


(b)

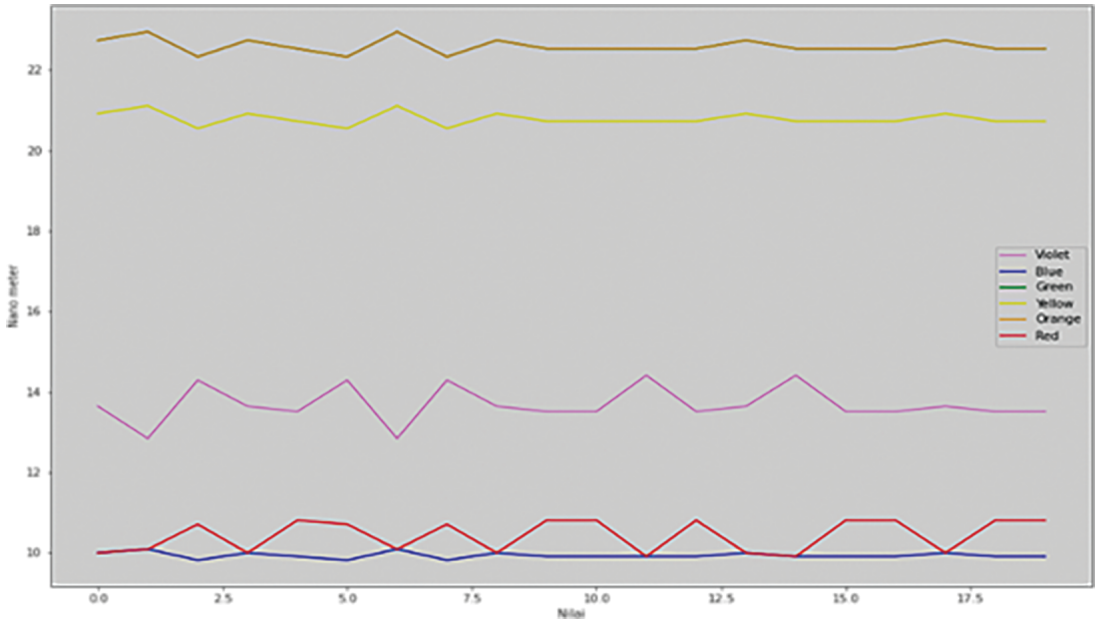


(c)

Figure 7: (Continued).



(d)



(e)

Figure 7: Graphs of the data results: (a) Normal, (b) Positive 1, (c) Positive 2, (d) Positive 3, and (e) Positive 4.

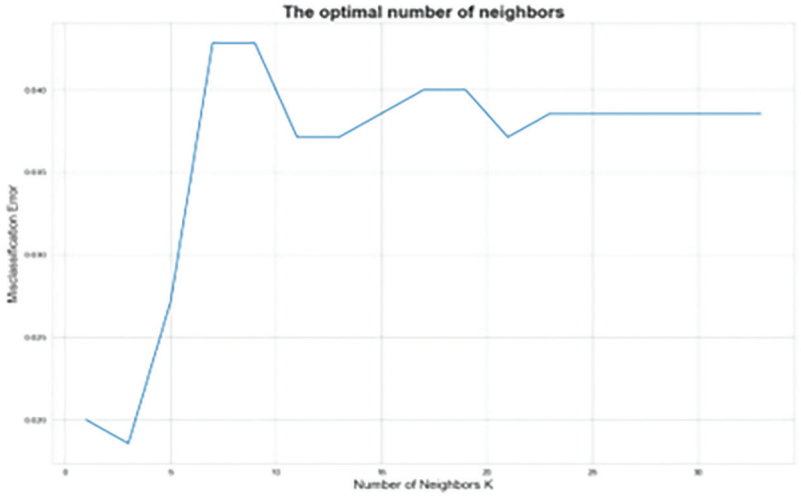


Figure 8: K-value graph.

	precision	recall	f1-score	support
0	0.97	0.98	0.98	61
1	1.00	0.96	0.98	56
2	0.95	0.95	0.95	60
3	0.90	1.00	0.95	57
5	1.00	0.92	0.96	66
accuracy			0.96	300
macro avg	0.96	0.96	0.96	300
weighted avg	0.97	0.96	0.96	300

Figure 9: KNN classification results. KNN, *K*-nearest neighbor.

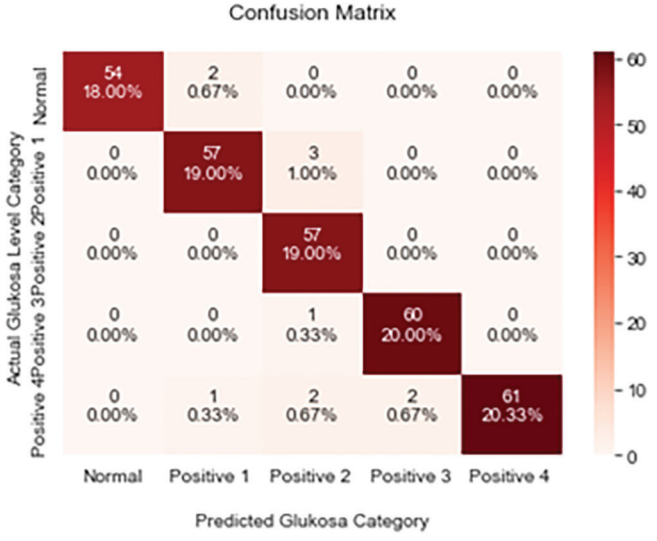


Figure 10: Confusion matrix.

Confusion matrix serves to provide information on the comparison of the classification results carried out by the system with the actual classification results [25].

4. Conclusions

The KNN algorithm can be quite effective in classifying the glucose class. In this study, each glucose level was determined by color matching based on the intensity of the colors violet, blue, green, orange, yellow, and red. The results obtained from the comparison made between the classification carried out by the system and the actual classification results showed a fairly large accuracy of 96.33%.

This system is designed to be used by many people, not necessarily diabetics because the examination is carried out without causing physical harm and other appliances do not come into direct contact with the body. In the future, we aim to create a noninvasive blood sugar–checking system with simpler hardware size and weight so that it is easier to carry anywhere.

Acknowledgments

This research was supported by The Ministry of Education, Culture, Research, and Technology of Indonesia, under contract number 002/PL.PTKN/BRIn.LPPM/VI/2022. We thank our colleagues from Universitas Ahmad Dahlan, who provided insight and expertise that greatly assisted the research.

References

[1] Kurniadi Wardana H, Indahwati E, Arifah Fitriyah L. Measurement of Non-Invasive Blood Glucose Level Based Sensor Color TCS3200 and Arduino. IOP Conf Ser Mater Sci Eng. 2018;336(1). doi:10.1088/1757-899X/336/1/012019

[2] Putri WJ, Isa I. Studi Literatur Glucose Sensor for Human Blood Using Glassy Carbon Electrode (GCE) Based Electrochemical Sensor with Voltammetry Method. J Fis Unand. 2021;10(3):324-329.

[3] Fridayanti N, Muldarisnur M. Design and Build a Blood Sugar Level Measuring Instrument in Urine with the Evanescent Method. Positron. 2018;8(2):1. doi:10.26418/positron.v8i2.26993

[4] Lengkong TD, Wowor MF, Berhimpon SLE. Overview of Blood Glucose and Urine Glucose in Young Adults Overweight and Obesity. Med Scope J. 2020;1(2):56-60. doi:10.35790/msj.1.2.2020.27816

[5] Febryansah, M. Iqbal, Anton Yudhana, and Alfian Ma'arif. 2020. "Smart Urinal Analyzer Detecting Abnormalities in Kidney Function With Analysis Of Ph Levels And Color In Urine." *Mobile and Forensics* 2(1): 32–40.

[6] Putra G. Design and Build a Non-Invasive Blood Sugar Level Measuring Device Based on Atmega 328P Microcontroller by Measuring the Turbidity Level of Urine Specimens Using Photodiode and Light Emiting Diode Sensor Packages. 2015;4.

[7] Satria, Eko W. Design and Build a Non-Invasive Blood Sugar Level Measuring Instrument Based on the AT89S51 Microcontroller by Measuring the Turbidity Level of Urine Specimens Using Photodiode Sensors. J Fis Unand. 2013;2(1):40-47.

[8] Permanasari, A. R. et al. 2018. "The Effect of Substrate and Enzyme Concentration on the Glucose Syrup Production from Red Sorghum Starch by Enzymatic Hydrolysis." IOP Conference Series: Earth and Environmental Science 160(1).[26]

[9] Pratiwi, Z., & Hufri. (2020). Manufacture of Blood Sugar Level Measuring Device Based on Turbidity Level of Urine Specimen Using Tcs230 Color Sensor And Photodiode With LCD Display. Pillar of Physics, 13(April), 18–25.

[10] Batta M. Machine Learning Algorithms - A Review. Int J Sci Res (IJ. 2020;9(1):381-undefined. doi:10.21275/ART20203995

[11] Leidiyana H. Application of the K-Nearest Neighbor Algorithm for Determining Credit Risk for Motor Vehicle Ownership. J Penelit Ilmu Komputer, Syst Embed Log. 2013;1(1):65-76.

[12] Anton Yudhana, et al. 2020. "Human Emotion Recognition Based on EEG Signal Using Fast Fourier Transform and K-Nearest Neighbor." *Advances in Science, Technology and Engineering Systems* 5(6): 1082–88.

[13] Ramadhan RS., Junta Z., Ardytha L. 2016. "Application of K-Nearest Neighbor Algorithm in Information Retrieval in Determining Final Project Reference Topics." 1(2): 123–33.

[14] Musliman, Anwar Siswanto, Abdul Fadlil, and Anton Yudhana. 2021. "Identification of White Blood Cells Using Machine Learning Classification Based on Feature Extraction." *Jurnal Online Informatika* 6(1): 63.

[15] Anton Yudhana, et al. 2022. "Fish Freshness Identification Using Digital Image-Based KNN Algorithm." 10(1): 1–9.

[16] Neighbor AK nearest, Chazar C, Nursyamsi I, Herwanto P. Breast Cancer Diagnosis Using Machine Learning With Introduction. 2021;(September): 182-191.

[17] Zhang S, Li X, Zong M, Zhu X, Cheng D. Learning k for kNN Classification. ACM Trans Intell Syst Technol. 2017;8(3). doi:10.1145/2990508

[18] Angreni IA, Adisasmita SA, Ramli MI, Hamid S. The Effect of K Value on the K-Nearest Neighbor (Knn) Method on the Accuracy Level of Road Damage Identification. Rekayasa Sipil. 2019;7(2):63. doi:10.22441/jrs.2018.v07.i2.01

[19] Anton Yudhana, Sunardi S, Hartanta AJS. K-Nn Algorithm With Euclidean Distance For Prediction Of Sengon Sawmill Results. Transmisi. 2020;22(4):123-129. doi:10.14710/transmisi.22.4.123-129

[20] Nikmatun IA, Waspada I. Implementation of Data Mining for Classification of Student Study Periods Using the K-Nearest Neighbor Algorithm. J SIMETRIS. 2019;10(2):421-432.

[21] Halim AAD, Anraeni S. Classification Analysis of Pneumonia Image Datasets using the K-Nearest Neighbor (KNN) Method. Indones J Data Sci. 2021;2(1):01-12. doi:10.33096/ijodas.v2i1.23

[22] Ente, Dewi Rahma et al. 2020. "Classification of Factors Causing Diabetes Mellitus in Unhas Hospital Using the C4.5. Algorithm." *Indonesian Journal of Statistics and Its Applications* 4(1): 80–88.

[23] Nikam SS. A Comparative Study of Classification Techniques in Data Mining Algorithms. Int J Mod Trends Eng Res. 2017;4(7):58-63. doi:10.21884/ijmter.2017.4211.vxayk

[24] Hutapea, Tanzil & Indriati. 2018. "Application of Modified K-Nearest Neighbor Algorithm in Classification of Schizophrenic Mental Illness." Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer 2(10): 3957–61.

[25] Tangkelayuk, Aldi. 2022. "The Water Quality Classification Using KNN, Naïve Bayes, And Decision Tree Methods." JATISI (Jurnal Teknik Informatika dan Sistem Informasi) 9(2):1109-19.

[26] Singh, Satyanand. 2020. "Minimal Redundancy Linear Array and Uniform Linear Arrays Beamforming Applications in 5G Smart Devices." *Emerging Science Journal* 4(Special Issue): 70–84.